

ADA 079214

LEVEL

①

RESEARCH MEMORANDUM 61-15

ESTABLISHING CUTTING SCORES FOR
ARMY LANGUAGE PROFICIENCY TESTS

OCTOBER 1961

DDC FILE COPY

DDC
RECEIVED
DEC 20 1979
A

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

79 12 18 294

Army Project Number
2195-60-001

Foreign Language a-41

⑨ m.e.m.o.
Research Memorandum 61-15
⑥ ESTABLISHING CUTTING SCORES FOR ARMY LANGUAGE
PROFICIENCY TESTS

⑩ R. G. Berkhouse, J. J. Mellinger K. Cook

⑭ AGO-HFRB-RM-61-15

Submitted by
E. Kenneth Karcher
Chief, Behavioral Evaluation Research Laboratory

⑪ Oct 1961

Accession For	
NTIS GFA&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	<input type="checkbox"/>
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or special
A	

This document is intended for use only by the Adjutant General, AGS, and is not available for distribution.

003 650

ESTABLISHING CUTTING SCORES FOR ARMY LANGUAGE PROFICIENCY TESTS

BACKGROUND AND PURPOSES

In 1954 Personnel Research Branch^{1/} research scientists, in coordination with the Army Language School, undertook to revise the Army Language Test (ALPT) which had been in use since 1949. These tests were designed to measure reading comprehension, aural comprehension, and writing ability of Army personnel assigned to linguistic jobs. The first step was the construction and field validation of revised language proficiency tests for two representative languages--Chinese-Mandarin and Russian. The final forms of these two tests were used as the prototypes for construction of 33 subsequent tests in other languages. It was assumed that by using this prototype approach the validity of the subsequent tests would be comparable to that of the two prototypes and that cutting scores on the subsequent tests could be generalized from those established for the two prototypes. The primary objective of this study was to investigate the comparability of validity coefficients and cutting scores of Army Language Proficiency Tests for three additional languages--French, German, and Polish--not previously analyzed.

PROCEDURES

EXPERIMENTAL TESTS

Three experimental tests described below were validated on Army personnel assigned to France and Germany.

1. Army Language Proficiency Test - French, DA PT 3439. This test is presented in two parts: Part I, Listening Comprehension and Part II, Reading Comprehension. Each part, containing 60 items, yields a maximum score of 60.

2. Army Language Proficiency Test - German, DA PT 3442, and

3. Army Language Proficiency Test - Polish, DA PT 3472 are identical to the French Test with respect to format and item types.

^{1/}On 5 December 1960, the organization was designated Personnel Research Branch, The Adjutant General's Office. Effective 1 January 1961 its title was changed to the present R and D Command facility.

In an earlier study (Dunn, T. F., et al. May 1957), Part I was found to relate highly with performance on Army type tasks involving conversational skills; a similar degree of relationship was established between Part II and performance on tasks involving reading and writing skills.

CRITERION MEASURES

Two expert linguists with fluency in both the appropriate foreign language (French, German, or Polish) and English evaluated each examinee's performance and translator work in simulated interpreter samples. Translator work samples (TWS) were administered and evaluated before the interpreter work sample (IWS). The work samples included:

Translator Work Samples -- French, DA PT 3638; German, DA PT 3640; Polish, DA PT 3642

Interpreter Work Samples -- French, DA PT 3639; German, DA PT 3641; Polish, DA PT 3643

The Translator Work Sample required an examinee to translate (write down) typed statements from the foreign language into English and vice versa. Fifteen statements were presented for each performance type. The translated statements were then evaluated by two experts. The total score was the sum of the evaluations of each statement. In most cases the same subject matter experts evaluated both translator and interpreter work sample performance. On the basis of findings in a previous study (Dunn, T. F., et al. May 1957), it was believed that no apparent influence obtains between work sample evaluations made by the same rater as compared with those made by different raters.

In the Interpreter Work Sample the examinee was required to perform as an interpreter between an English-speaking interrogator and a French-speaking informant. The examinee was required to listen to 30 statements or questions, 15 of which were read by the interrogator in English and restated by the examinee in the appropriate language. Then, 15 statements were given in the foreign language and required to be restated in English. Statements and questions used in the script made up an integrated conversation. Structure, format, and administrative procedures were identical for each of the three languages. The total score was the sum of the evaluations made by each of the language experts. An overall rating of the examinee's skill as an interpreter was prepared in each work sample on an 11-point rating scale covering five major skill levels of usefulness as an interpreter in terms of ability to communicate accurately and completely. An example of the scale is shown in Figure 1.

Studies of Army personnel who have been tested for their foreign language proficiency have shown a high relationship between reading

INTERPRETER RATING SKILLS

HOW USEFUL IS THIS MAN FOR AN INTERPRETING JOB IN TERMS OF HIS ABILITY TO COMMUNICATE?

- 31 _____ He could not handle any interpreter assignment.
- 32 _____
- 33 _____ He should be used as an interpreter ONLY in an emergency.
- 34 _____
- 35 _____
- 36 _____ He could be used on most lower level interpreter assignments.
- 37 _____
- 38 _____ He could be used on most interpreter assignments except those requiring highly literate native fluency or its equivalent.
- 39 _____
- 40 _____
- 41 _____ He would be successful at handling any kind of interpreter assignments.

Figure 1. Rating Scale - Interpreter Work Sample

comprehension skills and conversational usage skills. For example, in one study of the English Fluency Battery (Robinson, J. E., et al., April 1977) intercorrelation coefficients between speaking, understanding, and reading ability in English of 597 Puerto Ricans ranged from .50 to .75. In a study of the ALPT prototypes (Dunn, T. F., et al., May 1977) the range of intercorrelation coefficients between speaking-understanding type criteria (Interpreter-Audiomonitor Work Samples) and reading-writing type criteria (Translator Work Samples) was .59 to .78.

SAMPLES

The experimental tests were administered during May and June 1958 to 334 military personnel on duty in France and Germany. 116 examinees were given the ALPT-French; 115, the ALPT-German; and 103, the ALPT-Polish. Slightly less than one-half of the men were serving in jobs requiring proficiency in language skills at the time. Within this limited linguist segment (considered typical for Army linguist populations) a relatively broad range of language ability was evidenced by performance on the work samples. Table 1 shows score range, means, and standard deviations for each of the work samples for all three tests for this segment.

Table 1
SCORE RANGE, MEANS, AND STANDARD DEVIATIONS
ON LANGUAGE WORK SAMPLES FOR EXAMINEES ASSIGNED
TO LINGUISTIC JOBS

Language	Work Sample	Actual Score Range	Mean	S. D.
French	Translator	0 - 30 ^a	16.31	9.56
	Interpreter	0 - 60 ^b	39.43	18.50
German	Translator	0 - 30	16.52	7.34
	Interpreter	0 - 60	40.15	11.55
Polish	Translator	0 - 26	12.39	7.27
	Interpreter	0 - 60	38.02	14.55

^aThe possible range for the Translator Work Sample for all three languages was 0 - 30.

^bThe possible range for the Interpreter Work Sample for all three languages was 0 - 60.

ESTABLISHING CUTTING SCORES

To establish required levels of proficiency for the Army's needs, it was necessary to determine qualifying scores. Subject matter experts completed an item-by-item evaluation of an examinee's performance on the Interpreter Work Sample and also a rating of his overall performance. The rating scale contained "built-in" cutting points defined in terms of "Unsatisfactory", "Poor", "Fair", and "Good". Using the equal percentile method, criterion cutting could then be related to cutting scores on the language proficiency tests. Comparable ratings were not available for the translator work samples. However, on the basis of the range level of correlation coefficients between interpreter and translator work samples (.66 - .79), it was considered operationally feasible to use the Interpreter Work Sample cutting points for Translator Work Sample in determining cutting scores.

Operationally, both numerical and adjectival scores are recorded on Form 20, Soldier's Qualifying Record, for performance on the Army Language Proficiency Tests. The adjectival descriptions are "Good", "Fair", and "Poor". As has been mentioned, cutting points on the predictor measures were set by using the same percentile at which cutting points fell on the criterion measures. For example, it was determined that the bottom of the "Good" category fell at the 78th percentile on the rating form. This same percentile was used to set the cutting score for the bottom of the "Good" category on the predictor. For administrative purposes, it was desirable to establish a common cutting point at each of the descriptive levels for all language proficiency tests. In order to achieve one set of cutting scores on each part of the tests, the mean of all three language tests at each proficiency level was computed. The average cutting scores thus derived are given in Table 2.

Since it was also administratively desirable to use a common cutting score for 32 other language tests, it was important to have an indication of the amount of misclassification which would occur in using the generalization procedure. For this purpose an individual was considered to be misclassified if, as a result of using the common cutting score for all languages, he was placed in a different descriptive category (Good - Fair - Poor), than he would be if separate cutting scores were set for each language. Percentages were computed of cases misclassified (Table 2) in each category for each part of the three tests.

RESULTS

Interrelationships among the criterion and predictor variables as well as means and standard deviations for the French, German, and

Table 2

MEANS OF THE CUTTING SCORES FOR THREE LANGUAGES AND
PERCENT OF CASES WHICH WERE THUS MISCLASSIFIED AT EACH CATEGORY

Variable	Mean Cutting Score	Percent Misclassified		
		French	German	Polish
Listening Comprehension				
Good	48.5	0	6	10
Fair	37.5	1	2	13
Poor	21.5	4	10	3
Reading Comprehension				
Good	48.5	10	0	9
Fair	37.5	9	1	0
Poor	21.5	8	8	5

Polish Language Proficiency Tests are summarized in Table 3. In the current study, intercorrelation coefficients for the Translator Work Sample and the Interpreter Work Sample are: French .79, German .74, and Polish .66. These coefficients are consistent in magnitude with those found in other studies on the ALPT. Kuder-Richardson (formula 20) reliability coefficients for the two parts of each of the test were consistently high--.90 to .95. The specific values are reported in Table 3. For each of the two work samples on all three languages, Kuder-Richardson (formula 20) and inter-rater reliability coefficients were rather high (generally in the 90's). These estimates of criterion reliability are reported in Table 4. Analysis of Interpreter Work Sample scores in Part I against those for Translator Work Samples in Part II yielded validity coefficients of .83, .66, and .73, respectively for the French, German, and Polish tests.

Table 3
VALIDITY AND RELIABILITY COEFFICIENTS OF FRENCH, GERMAN, AND POLISH ARMY LANGUAGE PROFICIENCY TESTS

Variable	N	Maximum Possible Score	Mean	S. D.	Correlation with work samples Interpreter Translator	Kuder-Richardson Reliability (Formula 20)
FRENCH						
Listening Comprehension	116	60	39.2	14.6	.83	.95
Reading Comprehension	115	60	35.6	13.7	.86	.94
GERMAN						
Listening Comprehension	115	60	39.8	11.5	.66	.91
Reading Comprehension	115	60	41.5	12.0	.72	.93
POLISH						
Listening Comprehension	103	60	42.5	10.6	.73	.90
Reading Comprehension	103	60	39.5	12.8	.87	.93

Table 4

ESTIMATES OF RELIABILITY OF THE CRITERION MEASURES

Variable	Kuder-Richardson Reliability (Formula 21)	Inter-Rater Agreement
Translator Work Sample		
French	.94	.98
German	.89	.95
Polish	.90	.91
Interpreter Work Sample		
French	.95	.96
German	.95	.91
Polish	.90	.95

SUMMARY AND CONCLUSIONS

Based on the two prototypes, 33 additional language proficiency tests were constructed by staff members of the Army Language School in cooperation with research scientists of the Personnel Research Branch, The Adjutant General's Office. An attempt was made to adhere to prototype item content and composition to a sufficient extent to insure relative comparability in validity and difficulty for all of the new tests. The present Research Memorandum reports the results of the validation and statistical analysis undertaken for three tests not previously covered -- French, German, and Polish. Validity coefficients obtained were of sufficient magnitude to indicate that the three tests are highly efficient measures of language proficiency. Coefficients for these tests (.66 to .87) closely approximated those for the Chinese-Mandarin and Russian prototypes (.68 to .86). It was concluded that the tests were fairly comparable with respect to validity. Levels of fluency were designated "Good", "Fair", and "Poor" and cutting points were computed at these levels by use of equal percentiles on the criterion and on the predictor measures. A common set of cutting

↓ scores was established for both parts of the tests and for all three languages. Amounts of misclassification estimated to result from this procedure varied from 0% to 1% of the total cases. It was also concluded that validity and comparability assumptions had been met and that common cutting scores could be generalized to all tests of foreign language proficiency. ↗

REFERENCES

Publications of the Human Factors Research Branch, The Adjutant General's Research and Development Command

1. Robinson, John E., Rosenberg, Nathan, Kaplan, Harry, and Berkhouse, Rudolph C. On-The-Job Evaluation of the English Fluency Battery for Insular Puerto Ricans. Technical Research Report 1098. April 1957.
2. Dunn, Theodore F., Tye, Vermont M., Sternberg, Jack, and Berkhouse, Rudolph. Development and Evaluation of Prototype Army Language Proficiency Tests. Technical Research Report 1105. May 1957.